Classifying Music Genres Using Image Classification Neural Networks

Alan Kai Hassen, Hilko Janßen, Dennis Assenmacher, Mike Preuss and Igor Vatolkin

Abstract Domain tailored Convolutional Neural Networks (CNN) have been applied to music genre classification using spectrograms as visual audio representation. It is currently unclear whether domain tailored CNN architectures are superior to network architectures used in the field of image classification. This question arises, because image classification architectures have highly influenced the design of domain tailored network architectures. We examine, whether CNN architectures transferred from image classification are able to achieve similar performance compared to domain tailored CNN architectures used in genre classification. We compare domain tailored and image classification networks by testing their performance on two different datasets, the frequently used benchmarking dataset `GTZAN` and a newly created, much larger dataset. Our results show that the tested image classification network requires a significantly lower amount of resources and outperforms the domain specific network in our

Alan Kai Hassen · Hilko Janßen · Dennis Assenmacher · Mike Preuss
University of Münster, Information Systems and Statistics, Leonardo Campus 3, 48149 Münster
✉ alan.hassen@uni-muenster.de
✉ hilko.janssen@uni-muenster.de
✉ dennis.assenmacher@wi.uni-muenster.de
✉ mike.preuss@wi.uni-muenster.de

Igor Vatolkin
TU Dortmund, Department of Computer Science, Otto-Hahn-Str. 14, 44227 Dortmund
✉ igor.vatolkin@cs.tu-dortmund.de

Archives of Data Science, Series A
(Online First)
KIT Scientific Publishing
Vol. 5, No. 1, 2018

DOI: 10.5445/KSP/1000087327/20
ISSN 2363-9881

given settings, thus leading to the advantage that it is not necessary to spend expert efforts for the design of the network.

1 Introduction

The emergence of online audio platforms as Spotify, SoundCloud or Apple Music strengthens the need to describe audio files according to their characteristics. This makes them accessible for recommender systems, categorization or music discovery.

The scientific field of Music Information Retrieval deals with extracting information from audio files to solve a wide variety of problems, including artist identification, music recommendation and beat detection (Lamere, 2008). The extracted information can be used in music tags, which are high-level descriptions of the audio (Choi et al, 2016; e.g. the artist, emotion or year) and may be attached to tracks as labels. These music genre tags are the focus of this work.

Music genres, and their corresponding tags, are characterized by the common characteristics shared by their members. These are typically related to the instrumentation, rhythmic structure and harmonic content of the music (Tzanetakis and Cook, 2002). In the field of music genre classification, techniques from deep learning are emerging. These techniques (Zhang et al, 2016; Choi et al, 2016) which are defined as "computational models that are composed of multiple processing layers to learn representations of data with multiple levels of abstraction" (LeCun et al, 2015) have shown state-of-the-art improvements in comparison to traditional machine learning techniques in various fields as speech recognition, visual object recognition, object detection, drug discovery and genomics (LeCun et al, 2015).

Following this development, our work investigates the question how domain-tailored music genre classification networks perform compared to general image classification networks. As a domain-tailored network we consider a common network architecture, which has been adjusted in hopes of reaching a higher accuracy in a given domain. A finding of no significant difference between the networks' performances actually provides strong evidence for using state of the art image classification networks instead of designing domain tailored ones. As neural network design is a time consuming process, skipping the design phase

would save a lot of time and effort. Additionally, we introduce a new scientific dataset for benchmarking music genre classifiers.

The remainder of this paper is structured as follows: Section 2 deals with the relevant related work in the field of music genre classification and music information retrieval. In Section 3, the hypothesis of this work is introduced and explained in detail. The experimental setup employed to test the hypothesis is described in Section 4. We present the results of the experiments in Section 5 before we draw a conclusion of our work in Section 6.

2 CNN for Music and Image Classification

CNNs are neural networks designed to process data that is available in the form of multidimensional arrays (grids), as signals and sequences for one-dimensional arrays, images and audio spectrograms for two-dimensional arrays or video for three-dimensional arrays (LeCun et al, 2015).

In order to utilize CNNs, the audio data needs to be transformed into a visual representation (spectrogram) (Costa et al, 2017), for which several options are available and have been applied in the past: Short time fast Fourier transformation (STFT) (Zhang et al, 2016; Jeong and Lee, 2016; Rajanna et al, 2015), mel-spectrograms (Pons and Serra, 2017; Pons et al, 2017; Choi et al, 2016, 2017) and constant-Q transformation (CQT) spectrograms (Oramas et al, 2017). Using this visual representation, a CNN is applied assuming, that the network is able to detect auditory events in time-frequency representations by seeing them (Choi et al, 2016). According to Choi et al (2016) there are different reasons why the application of CNNs in the context of music genre classification might be beneficial.

First, that music tags, as for genres, are often considered as belonging to the most important high-level features representing track-level representation above intermediate-level features such as chords, beats, tonality and temporal envelopes which change over time and frequency. CNNs are designed to learn hierarchical features over a multilayer structure and are therefore suited to learn the hierarchy inherited in the genre classification task (Choi et al, 2016). Second, the properties of CNNs, in detail translation, distortion and local invariances, can be useful to detect relevant musical features that can appear at any time or frequency range (Choi et al, 2016).

There are several ways to include domain knowledge into a neural network design, like problem-adapted feature processing, design of specific architecture, or the setup of filter sizes for CNNs: In the domain of music information retrieval, three different conceptual filter sizes are applied in the convolutional layer: High filters, wide filters and small-rectangular filters (Pons et al, 2016). High filters use domain knowledge for designing filter shapes that can detect relevant time-frequency context in a spectrogram (Pons et al, 2017). High filters are used in the first convolutional layer only, with different filter sizes as 1x100 (width x height) for learning timbre (Pons et al, 2017) or alternatively to every convolutional layer in genre classification (Oramas et al, 2017). Related to the usage of high filters, wide filters have been suggested to find temporal dependencies within a spectrogram (Pons et al, 2016) and have found appliance in genre classification (Pons and Serra, 2017; Zhang et al, 2016), for example the usage of a filter size of 513x4 by Zhang et al (2016). Small-rectangular filters are applied by several studies without using domain knowledge for genre classification (Costa et al, 2017) and music tag classification (Choi et al, 2016, 2017), for example the usage of a filter size of 3x3 by Choi et al (2016). It is not entirely clear if domain specific filter designs are preferable over non-specific ones since high filters seem to perform better than small rectangular ones for genre classification (Oramas et al, 2017) and contrary, small rectangular filters seem to perform better than high filters for music tag classification (Choi et al, 2017).

Furthermore, there are various architecture options for CNNs available in the general domain of image classification, which is the task of producing a list of object categories present in an image (Russakovsky et al, 2015). An associated challenge is the Large Scale Visual Recognition Challenge (ILSVRC) (Russakovsky et al, 2015), functioning as a continuously improving benchmark for object category classification. The first deep CNN performing better than traditional techniques and winning in the ILSVRC 2012 was AlexNet (Krizhevsky et al, 2012), a CNN of depth 8. In the field of genre classification, parts of AlexNet have been adopted by the usage of a neural network architecture of 5x5 filter sizes, 2x2 max-pooling, a fully connected layer and a softmax layer (Costa et al, 2017).

The winner of ILSVRC 2014 VGG (Simonyan and Zisserman, 2014) employs deeper networks of 16 to 19 layers, consisting of convolutional layers with smaller filter sizes of 3x3 due to an increase in the discriminative abilities of the network and a decrease of network parameters. Those smaller filter sizes are

adapted in the field of music tagging by (Choi et al, 2016) by using a network of depth 9 with layers in total, consisting of four convolutional layers of filter size 3x3, four max pooling layers, a sigmoid function as a final layer and Rectified Linear Unit (ReLU; Nair and Hinton, 2010) as an activation function. ResNet (He et al, 2016a), won the ILSVRC 2015 by increasing the depth of a CNN to up to 152 layers and using residual connections over several layers, in the form of shortcut connections. The basic ResNet design was improved by changing the architecture of the inner residual block and consequently achieving better accuracies (He et al, 2016b). ResNet's concepts have been applied in genre classification with the usage of residual connections and a combination of max- and average pooling to build a neural network consisting of one residual block, max- and average-pooling, two fully connected layers and a softmax layer to predict music genres (Zhang et al, 2016). Newer network architectures in the field of image classification seem to perform better than ResNet, among them DenseNet (Huang et al, 2017). It connects each dense block, built out of a batch normalization layer, a ReLU activation function and a convolution layer of filter size 3x3, in a network with each other dense block in a feed-forward fashion. ResNext as used in (Xie et al, 2017) employs aggregations of multi-branch architectures within a residual block. Neither of those two networks architectures are, to our knowledge, applied in genre classification yet.

3 Hypothesis

In the field of genre classification, CNNs have been applied in several studies (compare Section 2). However, there is still the question at hand, if domain specific filters, applied as high (Oramas et al, 2017) and wide (Zhang et al, 2016) filters, and custom network architectures, applied by several researchers (Costa et al, 2017; Zhang et al, 2016; Choi et al, 2017), are superior to network architectures used in the field of image classification. This question arises, since image classification architectures have influenced the design of domain specific network architectures in, for example, the adaptation of parts from AlexNet (Krizhevsky et al, 2012) by Costa et al (2017), VGG (Simonyan and Zisserman, 2014) by Choi et al (2016) and ResNet (He et al, 2016a) by Zhang et al (2016).

Following this development and the reasoning for using CNNs for audio analysis tasks, we assume that the genre classification task, using spectrograms as audio representation and CNNs as a classifier, is an image classification task,

defined as producing a list of object categories present in an image (Russakovsky et al, 2015). We further assume, that the object categories of an image classification task are the genres of a track in the genre classification task.

Therefore, we hypothesize, that a CNN architecture applied in image classification is able to achieve comparable performance to a domain specific designed CNN architecture used in genre classification.

4 Experimental Setup

A classification problem can be seen as a coherent model selection and hyperparameter optimization problem, where the task is to find the right algorithm with the right hyperparameter settings for a dataset that optimizes the empirical performance (Thornton et al, 2013). In this work, two different audio datasets are used, namely `GTZAN` (Tzanetakis and Cook, 2002) and a newly created dataset, in the following called `10GenreGram`. In both cases the audio data is transformed into spectrograms, which serve as image representation input of the audio data for the neural networks. Furthermore, model selection and hyperparameter optimization are considered separately. This is common in the image classification domain (He et al, 2016a,b; Simonyan and Zisserman, 2014) and music genre classification domain (Zhang et al, 2016; Choi et al, 2016), which even omit the hyperparameter optimization task by choosing network parameters a priori. The models selected in the model selection are a domain-specific network (Zhang et al, 2016) and an image classification network (He et al, 2016a,b), since the possible performance differences between the two are relevant for this work. For both networks hyperparameters are optimized, k-fold cross validation (Kohavi, 1995) is applied, and the performance of both networks is compared.

4.1 Music Datasets

The first used dataset is `GTZAN` (Tzanetakis and Cook, 2002). It consists of 10 different genres: blues, classical, country, disco, hiphop, jazz, metal, pop, reggae and rock. Each of the genres consists of 100 tracks with 30 second snippets each, stored as 22050 Hz, 16-bit mono *.au audio files. Noteworthy, the dataset is noisy

with repetitions, mislabelings and distortions but it is used as the de facto standard dataset in the domain with more than 100 applications (Sturm, 2013). This allows for comparisons with other researchers operating at the same conditions.

Table 1: Genre distribution of both datasets used in our experiments.

	Genre				
Dataset	blues	classical	country	disco	hiphop
GTZAN	100	100	100	100	100
10GenreGram	1060	937	985	1079	1041
	jazz	metal	pop	reggae	rock
GTZAN	100	100	100	100	100
10GenreGram	1022	1018	976	1030	1005

Furthermore, we created an new dataset, the `10GenreGram` dataset, which consists of audio files downloaded from the music streaming platform SoundCloud (SoundCloud, 2018) and provides researchers with a higher magnitude of available data in comparison to `GTZAN` (see Table 1). For our experiment, we use only a subset, mirroring the 1000 tracks structure of `GTZAN` via subsampling due to expected long training times. The labeling of a track in the `10GenreGram` dataset is done by the uploader on the SoundCloud plattform by attaching descriptive tags to a track. These function as meta-data to describe the track by genres, moods or various other categories and can be used to find the most popular tracks for a tag. For the dataset, we use the most popular tracks from the following tags: Blues, classical, country, disco, hiphop, jazz, metal, pop, reggae and rock. The tags mirror GTZAN's genre structure and function as labels for our dataset which sets the `10GenreGram` dataset apart from for example the FMA dataset (Defferrard et al, 2017) which is not following the GTZAN genre structure. A drawback of labeling is the introduction of noise into the dataset because we trust the labeling of the track of its creator, assuming that it is in his best interest to label his tracks correctly to achieve high popularity. It should be emphasized that this approach results in the `10GenreGram` dataset being noisy just like `GTZAN`. However, this allows the creation of a much more extensive dataset. Furthermore, there is no guarantee that music genres are interpreted in the same way universally (Vlegels and Lievens, 2017).

After downloading the tracks, erroneous data and duplicates are removed. Erroneous data occurs when the track could not be downloaded, e.g. because of a SoundCloud restriction or an error in the SoundCloud downloader. After cleanup, the 10GenreGram dataset consists of 10,133 tracks with a total size of 69.93 GB and varying in length from 9 seconds to 6 hours and 31 minutes[1]. The audio properties are defined by SoundCloud with a sampling rate of 44100 Hz and 16 bit depth, stereo .mp3 audio files. Even though the goal was to gather 1000 tracks per genre, tracks are not equally distributed due to the removal of erroneous tracks and duplicates (see Table 1). This is neglectable since the deviation from the desired genre distribution is small.

4.2 Feature Extraction

We are utilizing the domain specific neural network of Zhang et al (2016) as a baseline for comparison. Therefore, we apply the same audio to spectogram procedure that they used within their work. Since the lengths of the tracks are not standardized for the 10GenreGram dataset and a transformation would, therefore, result in a non-uniform distribution of spectrogram files per audio file, the following strategy is applied to ensure an almost uniform distribution:

1. Only a 30-second or smaller window for each track is used to generate STFT spectrograms mirroring GTZAN's length. The design of the window follows the strategy by Costa et al (2017) which takes 60 seconds around the middle point of a track [−30 seconds, +30 seconds]. If a track is shorter than 60 seconds, the whole track is used.
2. Every track is split into 3-second windows following the practice of several researchers (Zhang et al, 2016; Pons et al, 2017; van den Oord et al, 2013) due to the fact that this has improved the accuracy while executing a majority voting system on track-level and has seen good results on GTZAN.
3. STFT transformation is performed using Librosa (McFee et al, 2015) with mono input and a sampling rate of 22050 Hz, following the GTZAN dataset properties and Librosa's limitation of only being able to create

[1] The dataset can be retrieved from https://10GenreGram.uni-muenster.de/.

mono spectrograms. Furthermore, a FTT window size of 1024 with Hann window function and hop size 512 is used.

4. The image size for every 3-second spectrogram is set to 128x513 pixels.

4.3 Classification Models

The two chosen networks are ResNet18 (He et al, 2016a) with improved residual block (He et al, 2016b) as an image classification network and NNet2 (Zhang et al, 2016) as a domain specific network. We have chosen these networks, since NNet2 is based on ResNet18 and can be seen as a good representative for a domain specific network. Additionally, it has been frequently used by other researchers.

Both networks are trained using 3-second spectrogram snippets and, therefore, predictions on track-level need to be generated to measure performance. This is achieved by using a majority vote approach (Zhang et al, 2016), where the probabilities of every 3-seconds spectrogram snippet from a track are added up and the genre with the maximum probability is chosen as the predicted label. Before usage, every spectrogram is normalized with zero mean and variance of one (Pons et al, 2017; Jeong and Lee, 2016). For all following experiments a batch size of 64 is used.

ResNet's optimal hyperparameters are chosen on `GTZAN` via tree-structured Parzen estimators (Bergstra et al, 2011) for 30 epochs and 50 evaluations with a hold-out split ratio of 80 % training, 10 % validation and 10 % testing on 3-second snippet-level without applying majority vote. The following parameters or components are optimized: Learning rate, momentum and weight decay of stochastic gradient descent (Hinton et al, 2012), furthermore the chosen activation function, ReLU or leaky ReLU (Maas et al, 2013), the L2 kernel regularizer penalty (Krogh and Hertz, 1991) and the chosen kernel initializer i.e. Xavier normal/uniform (Glorot and Bengio, 2010) or He normal/uniform (He et al, 2015).

For both networks, the found hyperparameter settings on `GTZAN` are then transfered to `10GenreGram`, assuming both datasets are comparable. Using the best found settings for both networks, a 10-fold cross-validation with 50 epochs per fold for both datasets is conducted to retrieve a more accurate estimate of

the networks' performance. Additionally, we use statistical testing to investigate whether the performance of both networks differs. It is important to note that the genres are balanced and that snippets of a track are assigned to exactly one fold. The performance comparison between both networks is done via a Wilcoxon singed-rank test (Wilcoxon, 1945) and a McNemar test with Edward's continuity correction (Edwards, 1948) on the fold results. Moreover, a comparison of the model complexity and the training duration is performed.

5 Results

In the following subsection we elaborate on the findings of our experiment. First, we present the results of our hyperparameter optimization. The accuracy and statistical testing are given in the following. We conclude by comparing the training times of both networks.

5.1 Hyperparameter Optimization

For the ResNet18, the hyperparameter optimization detected one combination of parameters which is superior to others: The stochastic gradient descent settings are a learning rate of 0.08, a momentum of 0.48 and a learning rate decay of 0.06. The kernels are initialized using the glorot normal initializer and furthermore the ReLU activation function is used. The L2 kernel regularizer uses a penalty value of 0.0004. With this combination of settings, a test accuracy of 77.1 % on snippet-level could be achieved. The NNet2 optimization showed that the given settings from Zhang et al (2016) performed best among the given alternatives. We could not increase the accuracy of the network choosing other settings.

5.2 Performance Comparison

For the GTZAN dataset, NNet2 reached an average accuracy of 72.3 % on snippet-level and an average accuracy of 80.4 % on track-level with majority vote in the 10-fold cross validation. In comparison, the hyperparameter optimized ResNet18 reached higher accuracies: On snippet-level, the achieved accuracy

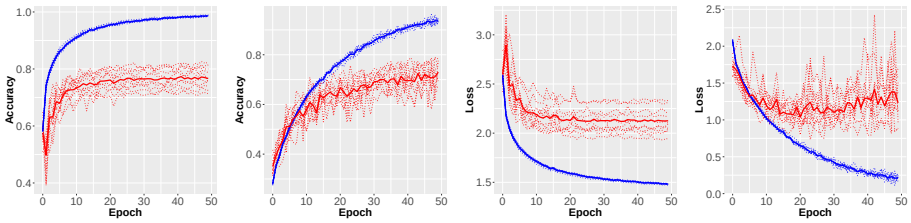
is 76.6 % and on track-level the accuracy is 84.7 %, averaging over all folds in both cases (see Table 2). A McNemar test with a critical alpha value of 0.05 and a p-value of $3.328 \cdot 10^{-4}$ shows that ResNet18 performed significantly better than NNet2. Likewise, the Wilcoxon signed-rank test also gets significant with a p-value of 0.01379. Comparing the training process of both networks, it is noticeable, that the ResNet reaches its final accuracy of the k-fold faster than NNet2 (see Figure 1). In the 10th epoch, ResNet accuracy is already close to its final accuracy after 50 epochs, while for NNet2, accuracy increases less steep. Furthermore, especially the loss variance of NNet2 is higher between different folds, compared to ResNet. The confusion matrices show a similar pattern for ResNet and NNet2. The only difference is the amount of misclassifications between both neural networks. NNet2 misclassifies the genres more often than ResNet, e.g. the misclassification rate for blues is 13 % for NNet2, but only 5 % for ResNet (compare Figure 3).

Table 2: Accuracies [%] for GTZAN dataset over 10 folds on track-level.

	Fold										
Network	1	2	3	4	5	6	7	8	9	10	**AVG**
NNet2	84	75	84	85	77	84	76	71	83	85	**80.4**
ResNet18	86	82	84	90	80	88	83	78	91	85	**84.7**

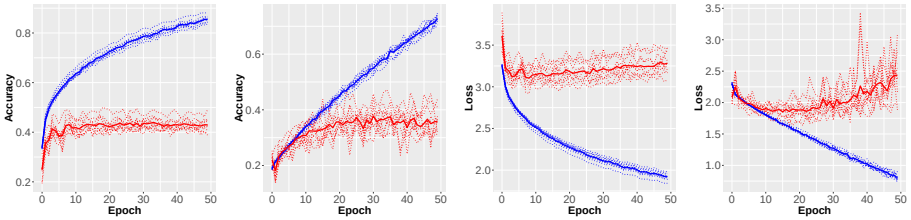
Table 3: Accuracies [%] for 10GenreGram dataset over 10 folds on track-level.

	Fold										
Network	1	2	3	4	5	6	7	8	9	10	**AVG**
NNet2	47	41	31	34	48	39	51	44	38	40	**41.3**
ResNet18	52	53	57	50	51	53	46	51	46	55	**51.4**



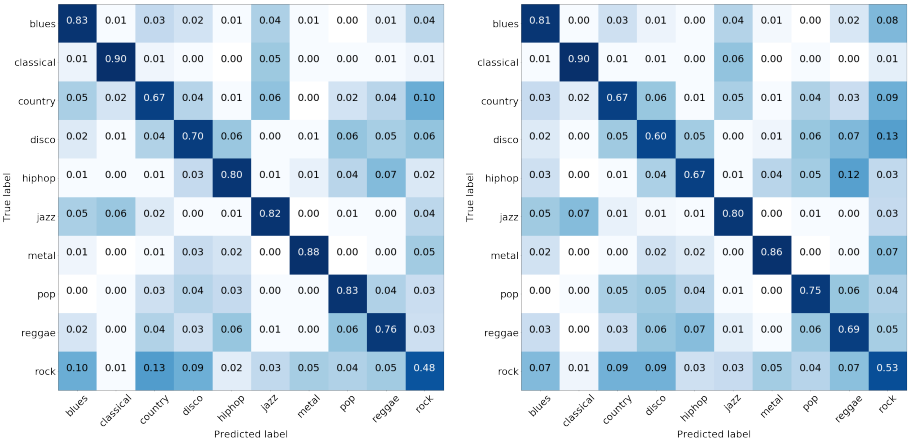
(a) ResNet Accuracy. (b) NNet2 Accuracy. (c) ResNet Loss. (d) NNet2 Loss.

Figure 1: GTZAN Training Graphs indicating training (blue) and validation (red) datasets.

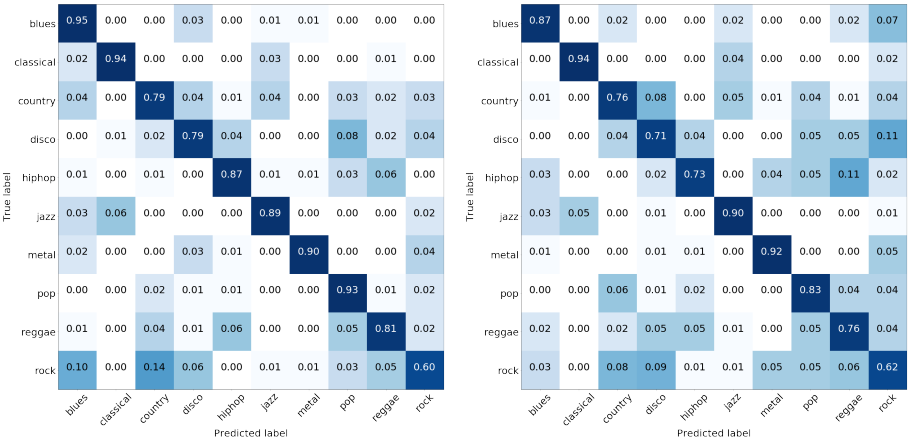


(a) ResNet Accuracy. (b) NNet2 Accuracy. (c) ResNet Loss. (d) NNet2 Loss.

Figure 2: 10GenreGram Training Graphs indicating training (blue) and validation (red) datasets.



(a) ResNet snippet-level. (b) NNet2 snippet-level.



(c) ResNet track-level. (d) NNet2 track-level.

Figure 3: Normalized GTZAN Confusion Matrix.

For the `10GenreGram` dataset, NNet2 achieved a 41.3 % accuracy on track-level with majority vote, while ResNet18 reached 51.4 % (see Table 3). The McNemar test for the `10GenreGram` dataset with a critical value of 0.05 and a p-value of $7.331 \cdot 10^{-9}$ shows a significant difference in performance between both classifiers. The Wilcoxon signed-rank test indicates a p-value of 0.01246 and thus confirms this. Concerning GPU memory usage, NNet2 requires a size of 32.66GB, which is a nearly six times bigger allocation then ResNet18 with 5.26GB. This implies, that for the training of NNet2 eight K80 are necessary, while for the training of the ResNet18 one K80 is enough. Even with more used GPUs, NNet2 needed significantly more time to train with 4.2 days compared to 4.4 hours with ResNet18.

6 Conclusion

Given our hypothesis, that an image classification network is able to achieve comparable performance to a domain specific designed CNN architecture, we can state that this is true. Not only do all conducted tests indicate a clear difference in classification performance, our image classification network additionally performed better on both datasets. Furthermore, the image classification network is less prone to overfitting, has a lower variance in the classification performance, is faster to train and needs lower hardware specifications doing so. Following those results it can be stated that using a classical image classification network is a viable alternative to domain specific networks in the field of genre classification with CNNs making the time consuming design phase of domain specific neural networks redundant. In the discussion whether domain specific filter sizes or small rectangular filter sizes are preferable (Oramas et al, 2017; Choi et al, 2017), this work indicates that networks using small rectangular filters with different levels of abstraction can deliver a better performance than networks using manually engineered domain specific filter sizes.

The limitations of this study are fourfold: First, for the `GTZAN` dataset, we were not fully able to reproduce the results given by Zhang et al (2016). According to the authors, they reached an accuracy of 87.4 %, which is 7 % higher than our reproduced results. Given the different error measurement practices, hold-out at Zhang et al (2016) and cross validation on our side, the different results might be explainable. Another factor can be that we had to reimplement and train the network from Zhang et al (2016) based on available

information, since no public implementation or model is available. Second, the results were achieved after training 50 epochs which can be a hindering factor if a network is not fully trained. It is important to note that the training graphs show a stagnant development which makes this unlikely. Third, we are aware that we are only comparing two neural networks and that our results cannot be generalized to all domain tailored neural networks. Lastly, we assumed that the uploader of the tracks for the `10GenreGram` dataset used the most suitable genre label. However, a manual expert labeling process could reduce the number of faulty labels.

References

Bergstra J, Bardenet R, Bengio Y, Kégl B (2011) Algorithms for Hyper-parameter Optimization. In: Proceedings of the 24th International Conference on Neural Information Processing Systems (NIPS'11), Shawe-Taylor J, Zemel R, Bartlett P, Pereira F, Weinberger K (eds), Curran Associates Inc., New York (USA), pp. 2546–2554. ISBN: 978-1-618395-99-3.

Choi K, Fazekas G, Sandler MB (2016) Automatic Tagging Using Deep Convolutional Neural Networks. In: Proceedings of the 17th International Society for Music Information Retrieval Conference (ISMIR 2016), Mandel MI, Devaney J, Turnbull D, Tzanetakis G (eds), pp. 805–811. ISBN: 978-0-692755-06-8, URL: `https://wp.nyu.edu/ismir2016/wp-content/uploads/sites/2294/2016/07/009_Paper.pdf`.

Choi K, Fazekas G, Sandler MB, Cho K (2017) Convolutional Recurrent Neural Networks for Music Classification. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2017), Institute of Electrical and Electronics Engineers (IEEE), New York (USA), pp. 2392–2396. DOI: 10.1109/ICASSP.2017.7952585.

Costa YMG, Oliveira LS, Jr. CNS (2017) An Evaluation of Convolutional Neural Networks for Music Classification Using Spectrograms. Applied Soft Computing 52:28–38. DOI: 10.1016/j.asoc.2016.12.024.

Defferrard M, Benzi K, Vandergheynst P, Bresson X (2017) FMA: A Dataset for Music Analysis. In: Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR 2017), Cunningham SJ, Duan Z, Hu X, Turnbull D (eds), pp. 316–323. ISBN: 978-9-811151-79-8, URL: `https://ismir2017.smcnus.org/wp-content/uploads/2017/10/75_Paper.pdf`.

Edwards AL (1948) Note on the "Correction for Continuity" in Testing the Significance of the Difference Between Correlated Proportions. Psychometrika 13(3):185–187. DOI: 10.1007/BF02289261.

Glorot X, Bengio Y (2010) Understanding the Difficulty of Training Deep Feedforward Neural Networks. In: Proceedings of the 13th International Conference on Artificial Intelligence and Statistics(AISTATS 2010), Teh YW, Titterington M (eds), Proceedings of Machine Learning Research (PMLR), Vol. 9, pp. 249–256. URL: http://proceedings.mlr.press/v9/glorot10a.html.

He K, Zhang X, Ren S, Sun J (2015) Delving Deep Into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In: 2015 IEEE International Conference on Computer Vision (ICCV 2015), IEEE Computer Society, New York (USA), pp. 1026–1034. DOI: 10.1109/ICCV.2015.123.

He K, Zhang X, Ren S, Sun J (2016a) Deep Residual Learning for Image Recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, (CVPR 2016), IEEE Computer Society, New York (USA), pp. 770–778. DOI: 10.1109/CVPR.2016.90.

He K, Zhang X, Ren S, Sun J (2016b) Identity Mappings in Deep Residual Networks. In: 14th European Conference on Computer Vision (ECCV 2016), Leibe B, Matas J, Sebe N, Welling M (eds), Springer, Cham (Switzerland), Lecture Notes in Computer Science (LNCS), Vol. 9908, pp. 630–645. DOI: 10.1007/978-3-319-46493-0_38.

Hinton G, Deng L, Yu D, Dahl G, rahman Mohamed A, Jaitly N, Senior A, Vanhoucke V, Nguyen P, Sainath T, Kingsbury B (2012) Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups. IEEE Signal Processing Magazine 29(6):82–97. DOI: 10.1109/msp.2012.2205597.

Huang G, Liu Z, van der Maaten L, Weinberger KQ (2017) Densely Connected Convolutional Networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, (CVPR 2017), IEEE Computer Society, New York (USA), pp. 2261–2269. DOI: 10.1109/CVPR.2017.243.

Jeong I, Lee K (2016) Learning Temporal Features Using a Deep Neural Network and its Application to Music Genre Classification. In: Proceedings of the 17th International Society for Music Information Retrieval Conference (ISMIR 2016), Mandel MI, Devaney J, Turnbull D, Tzanetakis G (eds), pp. 434–440. URL: https://wp.nyu.edu/ismir2016/wp-content/uploads/sites/2294/2016/07/159_Paper.pdf.

Kohavi R (1995) A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. In: Proceedings of the 14th International Joint Conference on Artificial intelligence (IJCAI'95), Volume 2, Morgan Kaufmann Publishers Inc., Montreal, Quebec, Canada (USA), pp. 1137–1143. ISBN: 978-1-558603-63-9, URL: http://dl.acm.org/citation.cfm?id=1643031.1643047.

Krizhevsky A, Sutskever I, Hinton GE (2012) ImageNet Classification with Deep Convolutional Neural Networks. In: Advances in Neural Information Processing Systems 25 (NIPS 2012): 26th Annual Conference on Neural Information Processing Systems 2012, Bartlett PL, Pereira FCN, Burges CJC, Bottou L, Weinberger KQ (eds), Neural Information Processing Systems Foundation, Inc., pp. 1097–1105. URL: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.

Krogh A, Hertz JA (1991) A Simple Weight Decay Can Improve Generalization. In: Advances in Neural Information Processing Systems 4, Moody JE, Hanson SJ, Lippmann R (eds), Morgan Kaufmann Publishers Inc., Burlington (USA), pp. 950–957. URL: http://papers.nips.cc/paper/563-a-simple-weight-decay-can-improve-generalization.

Lamere P (2008) Social Tagging and Music Information Retrieval. Journal of New Music Research 37(2):101–114. DOI: 10.1080/09298210802479284.

LeCun Y, Bengio Y, Hinton GE (2015) Deep learning. Nature 521(7553):436–444. DOI: 10.1038/nature14539.

Maas AL, Hannun AY, Ng AY (2013) Rectifier Nonlinearities Improve Neural Network Acoustic Models. In: Proceedings of ICML Workshop on Deep Learning for Audio, Speech, and Language Processing (ICML 2013), Vol. 28. URL: https://sites.google.com/site/deeplearningicml2013/accepted_papers.

McFee B, Raffel C, Liang D, Ellis DP, McVicar M, Battenberg E, Nieto O (2015) librosa: Audio and Music Signal Analysis in Python. In: Proceedings of the 14th Python in Science Conference (Scipy 2015), Huff K, Bergstra J (eds), pp. 18–25. DOI: 10.25080/Majora-7b98e3ed-003.

Nair V, Hinton GE (2010) Rectified Linear Units Improve Restricted Boltzmann Machines. In: Proceedings of the 27th International Conference on International Conference on Machine Learning (ICML'10), Omnipress, Madison (USA), pp. 807–814. ISBN: 978-1-605589-07-7, URL: https://dl.acm.org/doi/10.5555/3104322.3104425.

van den Oord A, Dieleman S, Schrauwen B (2013) Deep Content-based Music Recommendation. In: Advances in Neural Information Processing Systems 26 (NIPS 2013): 27th Annual Conference on Neural Information Processing Systems 2013, Burges CJC, Bottou L, Ghahramani Z, Weinberger KQ (eds), Neural Information Processing Systems Foundation, Inc., pp. 2643–2651. URL: http://papers.nips.cc/paper/5004-deep-content-based-music-recommendation.

Oramas S, Nieto O, Barbieri F, Serra X (2017) Multi-Label Music Genre Classification from Audio, Text and Images Using Deep Features. In: Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR 2017), Cunningham SJ, Duan Z, Hu X, Turnbull D (eds), pp. 23–30. URL: https://ismir2017.smcnus.org/wp-content/uploads/2017/10/126_Paper.pdf.

Pons J, Serra X (2017) Designing Efficient Architectures for Modeling Temporal Features with Convolutional Neural Networks. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2017), Institute of Electrical and Electronics Engineers (IEEE), New York (USA), pp. 2472–2476. DOI: 10.1109/ICASSP.2017.7952601.

Pons J, Lidy T, Serra X (2016) Experimenting with Musically Motivated Convolutional Neural Networks. In: 14th International Workshop on Content-Based Multimedia Indexing, (CBMI 2016), Institute of Electrical and Electronics Engineers (IEEE), New York (USA), pp. 1–6. DOI: 10.1109/CBMI.2016.7500246.

Pons J, Slizovskaia O, Gong R, Gómez E, Serra X (2017) Timbre Analysis of Music Audio Signals with Convolutional Neural Networks. In: 25th European Signal Processing Conference (EUSIPCO 2017), Institute of Electrical and Electronics Engineers (IEEE), New York (USA), pp. 2744–2748. DOI: 10.23919/EUSIPCO.2017.8081710.

Rajanna AR, Aryafar K, Shokoufandeh A, Ptucha R (2015) Deep Neural Networks: A Case Study for Music Genre Classification. In: 14th IEEE International Conference on Machine Learning and Applications (ICMLA 2015), Li T, Kurgan LA, Palade V, Goebel R, Holzinger A, Verspoor K, Wani MA (eds), Institute of Electrical and Electronics Engineers (IEEE), New York (USA), pp. 655–660. DOI: 10.1109/ICMLA.2015.160.

Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein MS, Berg AC, Li F (2015) ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision 115(3):211–252. DOI: 10.1007/s11263-015-0816-y.

Simonyan K, Zisserman A (2014) Very Deep Convolutional Networks for Large-Scale Image Recognition. CoRR. URL: http://arxiv.org/abs/1409.1556.

SoundCloud (2018) About SoundCloud – Listen to music. URL: https://soundcloud.com/pages/contact. Accessed on 5th January 2018.

Sturm BL (2013) The GTZAN dataset: Its Contents, Its Faults, Their Effects on Evaluation, and Its Future Use. CoRR. URL: http://arxiv.org/abs/1306.1461.

Thornton C, Hutter F, Hoos HH, Leyton-Brown K (2013) Auto-WEKA: Combined Selection and Hyperparameter Optimization of Classification Algorithms. In: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '13), Ghani R, Senator TE, Bradley P, Parekh R, He J (eds), Association for Computing Machinery (ACM), Chicago (USA), pp. 847–855. DOI: 10.1145/2487575.2487629, URL: http://dl.acm.org/citation.cfm?doid=2487575.2487629.

Tzanetakis G, Cook PR (2002) Musical Genre Classification of Audio Signals. IEEE Transactions on Speech and Audio Processing 10(5):293–302. DOI: 10.1109/TSA.2002.800560.

Vlegels J, Lievens J (2017) Music Classification, Genres, and Taste Patterns: A Ground-up Network Analysis on the Clustering of Artist Preferences. Poetics 60:76–89. DOI: 10.1016/j.poetic.2016.08.004.

Wilcoxon F (1945) Individual Comparisons by Ranking Methods. Biometrics Bulletin 1(6):80–83. DOI: 10.2307/3001968.

Xie S, Girshick RB, Dollár P, Tu Z, He K (2017) Aggregated Residual Transformations for Deep Neural Networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, (CVPR 2017), IEEE Computer Society, New York (USA), pp. 5987–5995. DOI: 10.1109/CVPR.2017.634.

Zhang W, Lei W, Xu X, Xing X (2016) Improved Music Genre Classification with Convolutional Neural Networks. In: 17th Annual Conference of the International Speech Communication Association (Interspeech 2016), Morgan N (ed), International Speech Communication Association (ISCA), pp. 3304–3308. DOI: 10.21437/Interspeech.2016-1236.